Project Report - Simulation Exercise

Course: Statistical Inference
Author: Shien Yang Lee

Overview

This report explores the behaviour and properties of the Central Limit Theorem (CLT) by applying it to an exponential distribution. 1000 independent samples of 40 observations each are simulated and their sample means calculated. A rate parameter, lambda, of 0.2 is held constant for all simulations carried out herein.

```
lambda = 0.2
```

Simulations

A single simulation entails drawing a sample of size 40 from the exponentially distributed population and calculating the sample mean. This procedure is repeated 1000 times to obtain a sample of 1000 empirical means, which is stored in the vector 'xbar'. Subsequent analysis involving the CLT will focus on this collection of sample means.

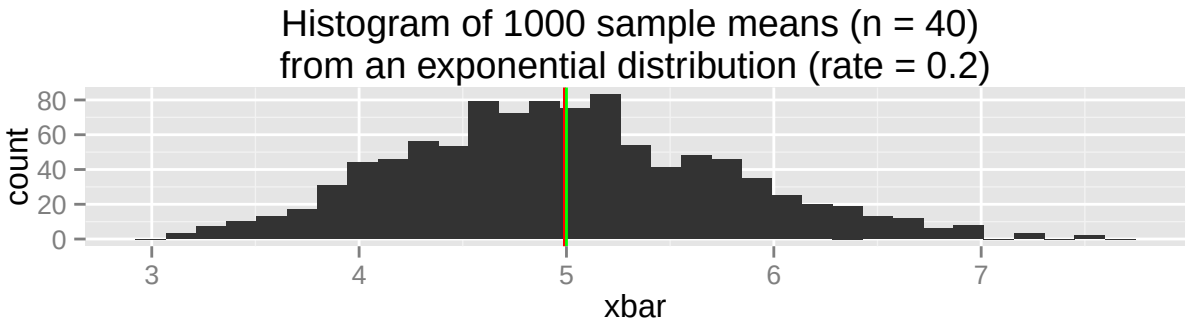
```
set.seed(1)
xbar = NULL
x = NULL
for (i in 1:1000) {
  sample = rexp(40,rate = lambda)
  x = append(x, sample, length(x))
  xbar[i] = mean(sample)
}
```

Sample Mean vs. Theoretical Mean

Since the sample mean is an unbiased estimate for population mean, the expected value of the sample mean is equal to the true population mean of the underlying distribution. In this case, we know that the underlying population is exponentially distributed with a theoretical mean of 5.

```
E_xbar = mean(xbar)
mean_prct_err = (E_xbar - (1/lambda))/(1/lambda) * 100

library(ggplot2)
mean_plot = ggplot(data = data.frame(xbar = xbar),
                   aes(x = xbar))
mean_plot = mean_plot + geom_histogram()
mean_plot = mean_plot + geom_vline(xintercept = E_xbar,
                                   colour = "red")
mean_plot = mean_plot + geom_vline(xintercept = 1/lambda,
                                   colour = "green")
mean_plot = mean_plot + ggtitle(paste("Histogram of 1000 sample means (n = 40)",
                                      "\nfrom an exponential distribution (rate = 0.2)"))
print(mean_plot)
```



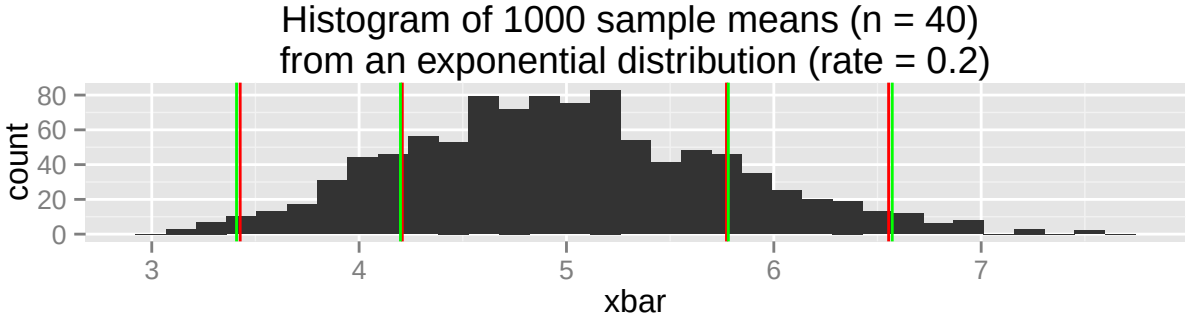
The empirical mean of the 1000 sample means is `4.9900252`, which represents a `-0.199496`% difference from the theoretical mean of `5`. The empirical mean (red line) and theoretical mean (green line) are plotted on a histogram of the 1000 sample means, showing the close agreement between the two values.

Sample Variance vs. Theoretical Variance

Another important result underpinning the postulates of the CLT is that the standard deviation of the sample mean, commonly known as the standard error of the mean, is related to the standard deviation of the underlying distribution by a factor of 1/(n^0.5). In this analysis, the sample and theoretical variability will be evaluated by comparing variances instead of standard deviations.

```
Var_hat_xbar = (sd(xbar))^2
Var_x = (1/lambda)^2
Var_xbar = Var_x/40
Var_prct_err = (Var_hat_xbar - Var_xbar)/Var_xbar * 100

library(ggplot2)
mean_plot = ggplot(data = data.frame(xbar = xbar),
                   aes(x = xbar))
mean_plot = mean_plot + geom_histogram()
mean_plot = mean_plot + geom_vline(xintercept = c(-2,-1,1,2) * sqrt(Var_hat_xbar) + E_xbar,
                                   colour = "red")
mean_plot = mean_plot + geom_vline(xintercept = c(-2,-1,1,2) * sqrt(Var_xbar) + E_xbar,
                                   colour = "green")
mean_plot = mean_plot + ggtitle(paste("Histogram of 1000 sample means (n = 40)",
                                  "\nfrom an exponential distribution (rate = 0.2)"))
print(mean_plot)
```



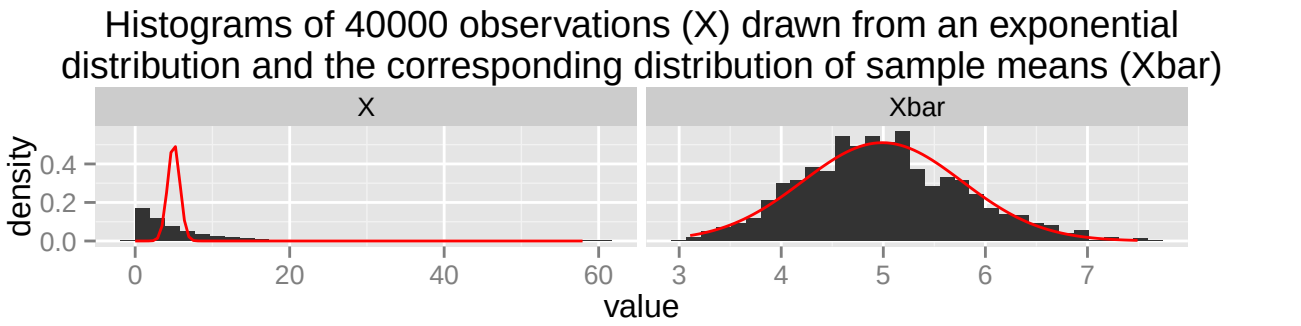
The empirical variance of the 1000 sample means is `0.6111165`, which represents a `-2.2213654`% difference from the theoretical variance of `0.625`. The red and green lines superimposed on the histogram represent the quantiles corresponding to 1 and 2 standard deviations away from the empirical mean. As expected from the negative value of the percentage error, the sample standard deviation (and thus sample variance) is slightly smaller than the theoretical variance while generally exhibiting close agreement.

Distribution

In addition to specifying the mean and standard deviation of the distribution of sample means, the CLT goes further to state that the distribution of sample means approach normality for large sample sizes.

```
variable = append(seq(1,1,length=length(x)),seq(0,0,length=length(xbar)),after=length(x))
variable = factor(x = variable, levels = c(1,0), labels = c("X","Xbar"))
df = data.frame(value = append(x,xbar,after=length(x)), variable=variable)

library(ggplot2)
distr_plot = ggplot(data = df, aes(x = value))
distr_plot = distr_plot + geom_histogram(mapping=aes(y = ..density..))
distr_plot = distr_plot + stat_function(fun=dnorm, colour = "red",
                                        args = list(mean = mean(xbar), sd = sd(xbar)))
distr_plot = distr_plot + facet_grid(.~variable, scales ="free")
distr_plot = distr_plot + ggtitle(paste("Histograms of 40000 observations (X) drawn from",
                                        "an exponential\ndistribution and the corresponding",
                                        "distribution of sample means (Xbar)"))
print(distr_plot)
```



The probability density function (red curve) of the normal distribution predicted by the CLT is superimposed on density histograms of the 40000 observations from the exponential distribution, as well as the collection of 1000 sample means. It is clear from the plot that the sample means approximate the normal distribution predicted by the CLT closely despite the fact that the underlying exponential data clearly falls under a non-normal distribution.